Data Integration and VISualization (DIVIS) : from large heterogeneous datasets to interpretable visualizations in plant science.

O. Thierry[1], R. Boumaza[1], J. Buitink[1], C. Landès[1], O. Leprince[1], M. Orsel[1] , P. Santagostini[1] et J. Bourbeillon[1]

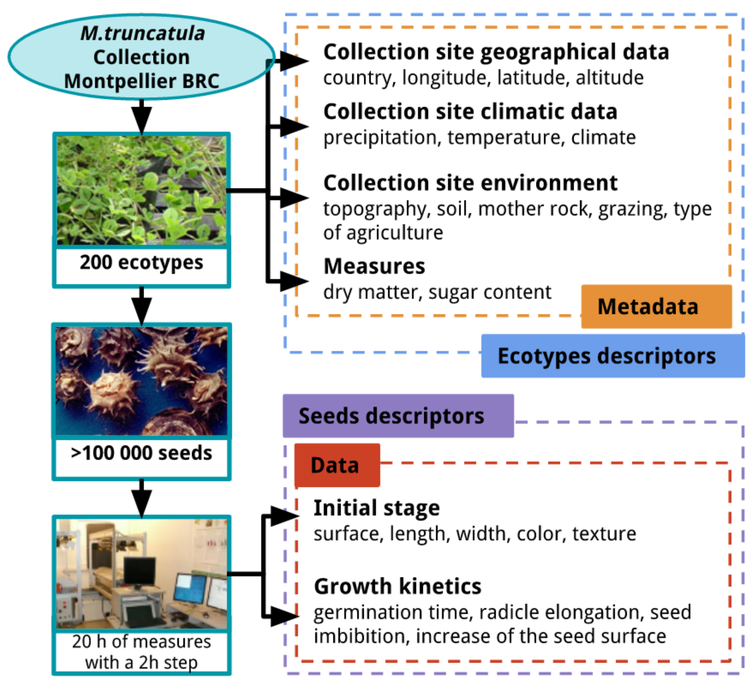
[1] *IRHS, Agrocampus-Ouest, INRA, Université d'Angers, SFR 4207 QuaSaV 42 rue Georges Morel, 49071 Beaucouzé Cedex, France*

Introduction

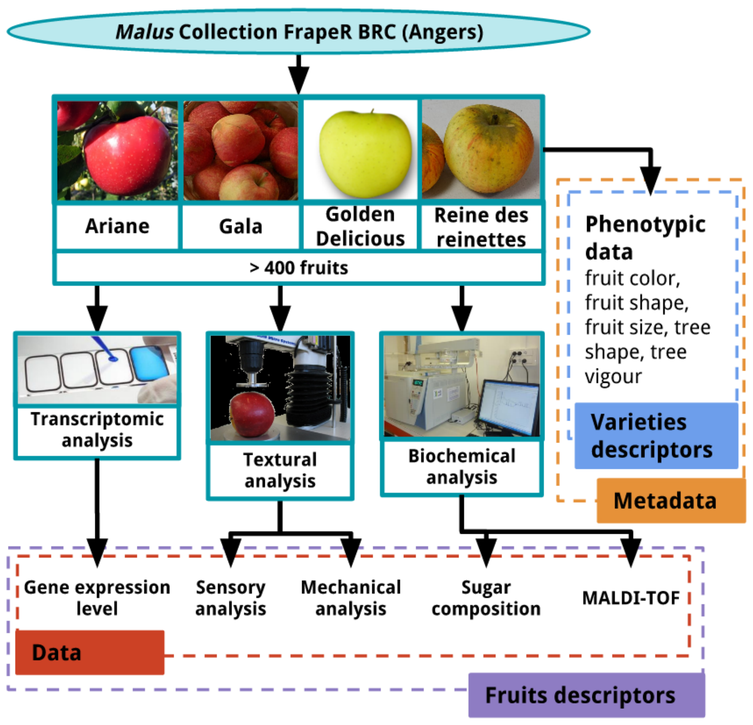
- The increase in the scale of experiments, the multiplication of experimental techniques used as part of one study and the generalized storage of past experimental datasets have led to the creation of large collections of heterogeneous datasets such as those stored by the IRHS teams in plant science.
- Currently, biologists are more and more interested in reusing and exploring these collections through integration and visualization techniques, but there is no user friendly tool addressing the whole problem, in particular in plant science [1-3].
- The DIVIS project aims to tackle the issue through a methodological exploration of a possible solution and the development of a directly usable prototype software on two test datasets, by combining the most promising integration and visualization approaches that are publicly available.

Datasets

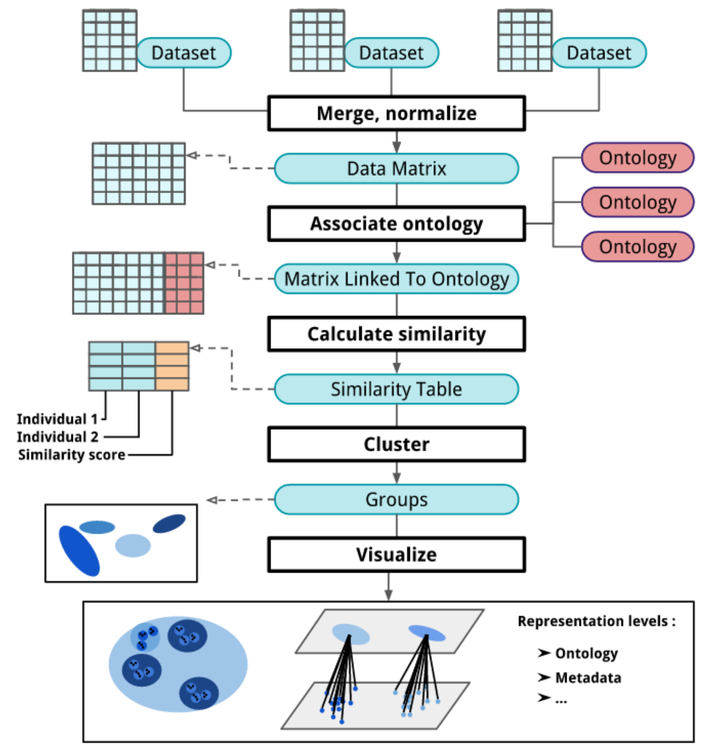
• Seed dataset



• Apple dataset



Software Workflow

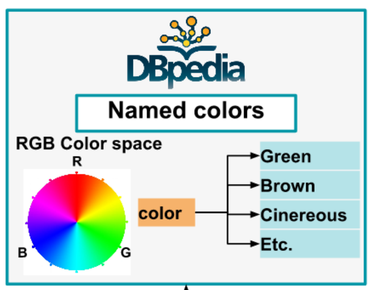


Project status

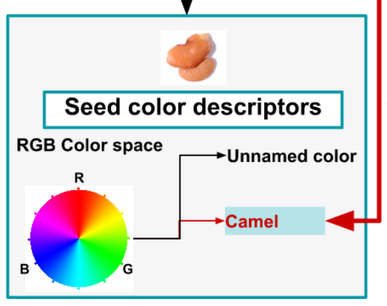
- Built the data matrices for both datasets.
- Performed basic statistical analyses to better understand the datasets.
- Performed some test visualizations with crude grouping approaches.
- Identified relevant existing ontologies and started designing missing ones.
- Current stage: associate metadata with ontologies and estimate their distribution [4-7].

Seed dataset: ontologies

• Color

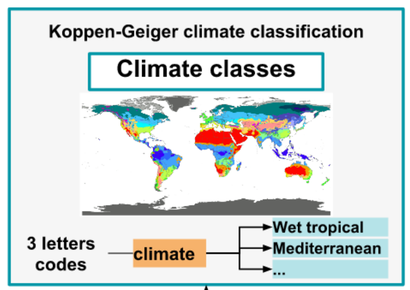


Distance

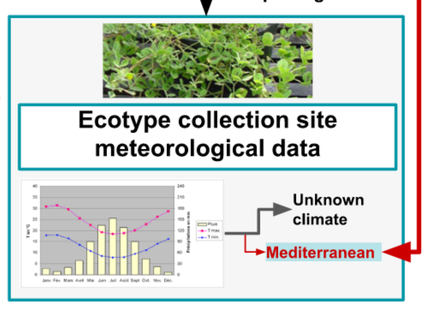


Attribute Named Colors

• Climate

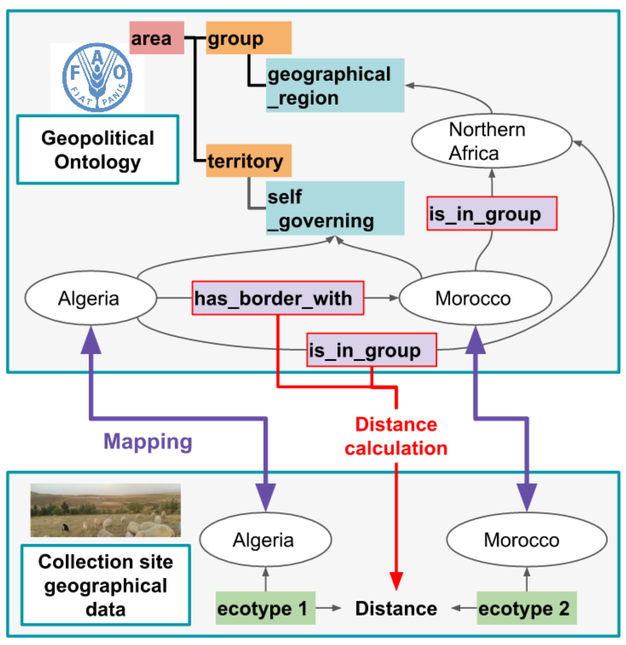


ClimClass R package

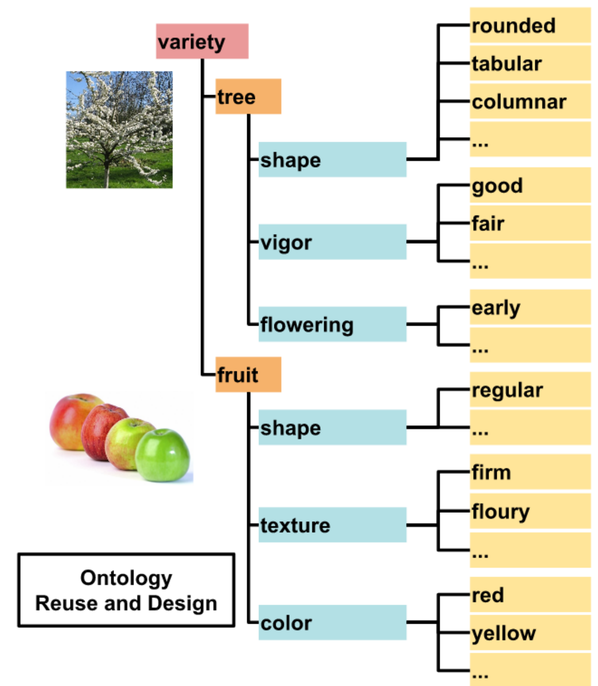


Attribute Climate class

• Country



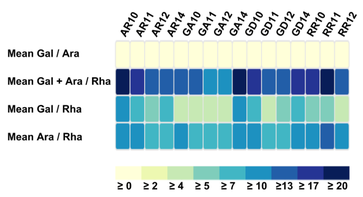
Apple dataset: ontologies



Visualization attempts

• Apple

Sugar contents



Organization axes

- mean sugar ratio
- sample

• Seed

Seed count plasticity

Organization axes

- region
- country
- precipitations class
- ecotype

▶ Biologists show interest in the displays.

Limits

- Need to easily navigate in the variables and individuals spaces.
  ▶ More interactive visualization.

Conclusion

- Preliminary operations in order to create a tool able to efficiently summarize and visualize larges heterogeneous datasets.
- Limited variability for the current metadata variables in the test datasets.
  ▶ Exploring other dimensions, such as shapes of seeds, might help to generate more different groups.
- Current visualization attempts performed using manually organized data files and as such predefined organization axes.
  ▶ A strong need to select and order the organization axes, that is to say the metadata variables to consider, emerged, which is considered in the final tool.
  ▶ Beyond the supervised approach, unsupervised clustering approaches might be of interest to drive the hierarchical organization of the individuals.

References

[1] Tao, S., Cui, L., Wu, X., Zhang, G.Q. (2017) Facilitating Cohort Discovery by Enhancing Ontology Exploration, Query Management and Query Sharing for Large Clinical Data Repositories. *AMIA Annu Symp Proc* 2017: 1685-1694.

[2] Solovieva, E., Shikanai, T., Fujita N., Narimatsu, H. (2018) GGDonto ontology as a knowledge-base for genetic diseases and disorders of glycan metabolism and their causative genes. *Journal of Biomedical Semantics* 9(14).

[3] Hendler, J. (2014) Data Integration for Heterogenous Datasets. *Big Data* 2(4):205-215.

[4] Fonseca, F.T., Egenhofer, M. J., Agouris, P. and Câmara, G. (2002) Using Ontologies for Integrated Geographic Information Systems. *Transactions in GIS* 6: 231-257.

[5] Kohler, S. Improved ontology-based similarity calculations using a study-wise annotation model.(2018) *Database: The Journal of Biological Databases and Curation* 2018:bay026.

[6] Cohen J., Matthen M. Color Ontology and Color Science. *Bradford Book, A.* (2010).

[7] Hartmann, J., Palma, R., Gómez-Pérez, A. Ontology Repositories. Handbook on Ontologies. Berlin, Heidelberg. Springer Berlin Heidelberg. (2009)

Acknowledgements

This work was partially funded by the projects: REGULEG ANR-15-CE20-000, AI-fruits ("Région Pays de la Loire") and DIVIS (RFI "Objectif Végétal"). We would like to thank the ANAN (transcriptomic analysis) and PHENOTIC (seed germination imaging) platforms for their role in the datasets acquisition.